e-ISSN : 2620 3502
p-ISSN : 2615 3785

Speech Recognition: Transcription and transformation of human speech

Vishal Dineshkumar Soni[1]

[1] *Department of Information Technology, Campbellsville University, Campbellsville, Kentucky*
Email: vishaldksoni@gmail.com

ABSTRACT

The specified subfield of computational linguistics and computer science can said to be linked with speech recognition. Speech recognition can develop new variation technologies as well as methodologies generated as interdisciplinary concept. It can be considered to translate and recognize and satisfy the capability towards understanding and translating the words that are already spoken. It is more preciously said that in the most recent times this field has secured positive feedback by intense learning of voice recognition. Such evidences shows the proof that it has more market demand for implementing the application of specific data as voice recognition. Deployment of speech recognition systems can be utilized as the evidence shown to its analyzing methods that is helpful for designing each and every individual's future. It is said that the computer plays an important role for this process as by this all the translated words can be acknowledged by the texts also.

Keywords: Speech recognition, computational linguistics, transcription, demotic appliance control, automated speech recognition.

1. INTRODUCTION

Speech recognition concept can be suggested for new variation of technologies as well as methodologies generated as interdisciplinary concept. Call home or voice dialing and call routing can be determined under demotic appliance control. It can be suggested for connecting an automatic call and also can be required for searching basic key words. Simplified data entry, searching specific podcast to say specific words, to enter the vital credit card numbers, analyzing and preparing structured documents as a kind of radiology report all the following above initiates determining the features of the speaker. When the requirement is seen to be raised as arranging some speech to text format as emails or the word processors then the direct voice input speech recognition applications are required for initial procedure.

2. OVERVIEW

ASR or automated speech recognition can be recognized as speech to text 'STT' or computer speech recognition. Computer engineering fields, linguistics, computer science fields all are responsible for incorporating the knowledge and research that are linked with speech recognition. Enrollment can be defined in the terms of speech recognition as the systems that are initially initiate the procedure of 'training' (Hsu et al. 2018). It can be stated as the system where the only speaker as an individual speaks out or read out the text. It can disseminate the systematic norm of isolated vocabulary required for the system. An individual's particular voice can be analyzed by the system. But it can be recognized in the perfection of tone that serves the capability to recognize the speech of those particular individual.Speaker independent systems can be recognized as the systems where the training doesn't matter or not required. Sometimes, this kind of terminology as direct voice input speech recognition applications can be assumed for voice user interfaces. Identification of the speaker voice can be recognized by their tone of the speech that can be referred to the speaker identification even the voice recognition. It is essential for the security process that it should be initiated to recognize the speaker voice. It can be fruitful towards simplifying the work as it can be translated according to the speech translation system.

Recognized speech can be enacted as a source towards authenticating and keeping the confidentiality of the information. Some kind of training can also be given to the particular person for this purpose that can be initially said as the voice recognition training. Speech recognition can serve an elongated history if it is generalized from the technological prospective. Voice recognition can be defined as the system where only the speaker as an individual can speak out read out the text and spread the specific information. It can disseminate the systematic norm of isolated vocabulary required for the system (Iter et al. 2017). An individual's particular voice can be analyzed by the system. Lots of huge variations of innovations are mandated for implementing such technique of voice recognition.

3. HISTORY OF SPEECH RECOGNITION

Speaker independence, processing speed and vocabulary size can be determined as the structure of suggestive key segments from where the development can be initiated.

1958 **–** R. Biddulph, K. H. Davis and Stephen Balashek as a three Bell Labs researchers were subjected to construct a system stated as Single-speaker digit recognition recommended for Audrey. In each of the utterance the power spectrum can be segmented as the system located in the formants.

1960 – The source-filter model originated from speech production can be published and developed by 'Gunnar Fant'**.**

1962 – In 1962, at World's Fair, the IBM can be subjected to demonstrate its 16-word "Shoebox" that has the capacity to handle machine's involuntary speech recognition system (Hsu et al. 2017).

1966 – A speech coding method can be conducted as (LPC) or linear predictive coding. Fumitada Itakura student of the University of Nagoya has proposed it for the first time. When it was in the form to work for speech recognition few person of (NTT) as Nippon Telegraph and Telephone named Shuzo Saito has shown its performance on speech coding method.

1969 – In 1969, the funding at Bell Labs was in dormant condition for many years. Again, when an open letter was shown by the effective John Pierce then the analytical analysis is served as defunded research for speech recognition. Until the retirement of Pierce, the specific kind of defunding stayed. After him, James L. Flanagan has taken his ascendant.

Constant speech was delivered by Raj Reddy on speech recognition for the first time. In the late 1960s, he was graduated from the University of Stanford. Pause after each of the words is essentially required by the previous systems. Playing chess has also generated a specific spoken command for playing by Reddy's system by that time.

By the specific time period, the dynamic time warping (DTW) algorithm served its invention by its soviet researchers. 200 per words vocabulary created by the researcher meant to be capable by the recognizer for serving the operations. Short frames are allowed for the processing of DTW that generates speech. Single unit for processing speech allows 10ms segments in its short frames (Schönherr et al. 2018). Though, with the help of later algorithms, the DTW are said to be superseded and the technique is processed according to dynamic time warping. According to the specific time period, achievement of the independence by its speaker states its critical analysis over dynamic time warping.

4. SPEECH RECOGNITION ON PRACTICAL BASIS

The n-gram language model was introduced in the year 1980 that show its efficiency towards speech recognition. Multiple length n-grams as the language models can be utilized by the back-off model. It serves its allowance towards recognizing languages as CSELT and HMM utilized in both the specialized processors of hardware and software. Example can be taken as 'RIPAC'.

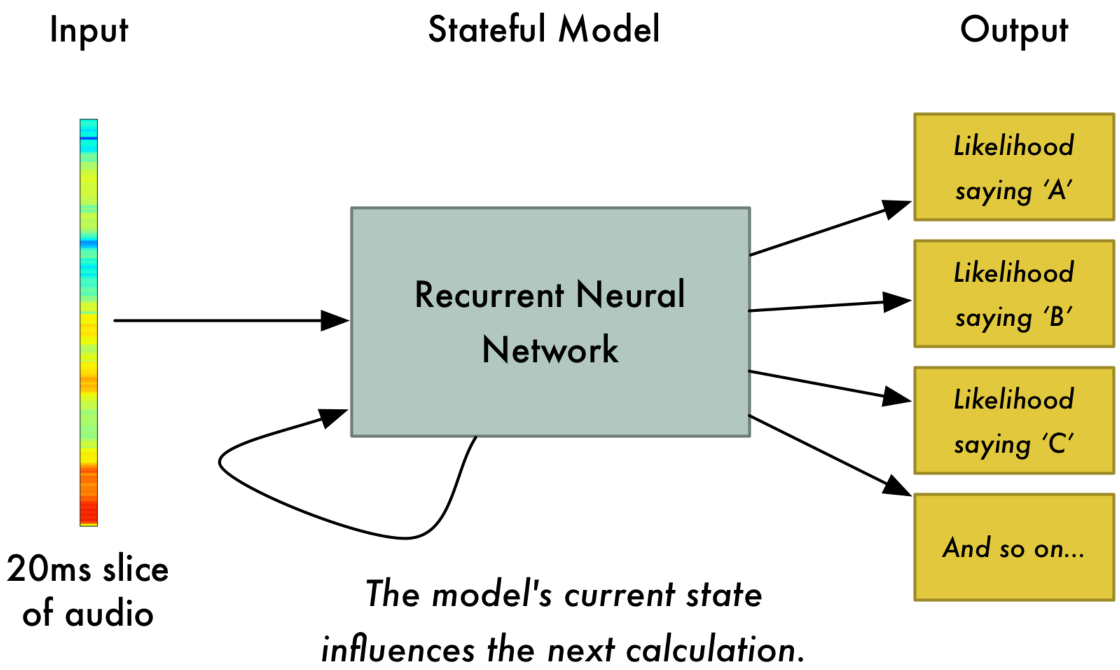

Figure 1: Speech recognition

(Source: Barker et al. 2017, p.641)

The intensity of capabilities is signified to be raised and seen in computers and the determined progress reports serves its positive value in this specified field (Fontan et al. 2017). In early 1976, at the ending stage of the DARPA program, it was initiated to develop PDP-10 along with 4 MB ram by the researchers as the best computer. 30 seconds of uttered speech can be decoded within 100 minutes via this computer. Practical products of two kinds are shown as:

- **1987** –Kurzweil Applied Intelligence has initiated its recognizer to develop practical products.
- **1990** – In early 1990, a consumer product was probably initiated to be released by Dragon Dictate. Without any kind of utilization offered by the human operator AT&T, it can insist the base of deployment on the basis of Voice Recognition Call Processing services. It can serve theroot canals on its telephonic calls. At Bell Labs, Lawrence Rabiner and many others developed this kind of technology.

Voice recognition was clearly evolved by speaker recognition since 2010s prospective of speech recognition. Major kind of breakthrough was seen after the independency of the speaker. 'Training' period is required for system analysis until then. The tagline carried by the doll can be initiated in 1987 that can be stated as the doll who can easily understand the speech of the person. In spite of that it can be subjected as the trainer to be suggestive for the children who can train their voices according to the gained response.

Milestone of a historical human parity can be reached through the specific Microsoft researchers. Transcribing can be initiated as the conversational telephonic speech that can be assumed on Switchboard task which is benchmarked in its early 2017 (Salimbajevs and Ikauniece 2017). Various intense learning models were utilized for optimization of speech recognizing perfection. All the professional human transcribers rated per word error rate as low as 4 being its essentials. On the same task, all the speech recognized report was said to be funded by IBM Watson under its same benchmark.

5. SPEECH RECOGNITION CAN BE BASED ON DYNAMIC TIME WARPING (DTW)

Speech recognition utilizes this approach of dynamic time warping to be signified since the period of its recognized histories. Since that time period the HMM-based approach is successfully working. Time or speed invests two sequences for which dynamic time warping can be signified as an algorithm. It measures the specific similarities. Observation can be detected by the person's movement. Instantly if the walking pattern of an individual is detected in a video that initiates accelerations and deceleration occurs due to the fast and slow movement analyzed by the walking patterns of a person. In the video, the whole procedure can be seen in a single clip without any blocking or we can say that the whole procedure takes place in a cut. Audio, video and graphics are applicable with the systematic pattern of DTW.DTW analyzes specific kind of linear representation that was the initial form of raw confirmed data (Tjandra et al. 2017).

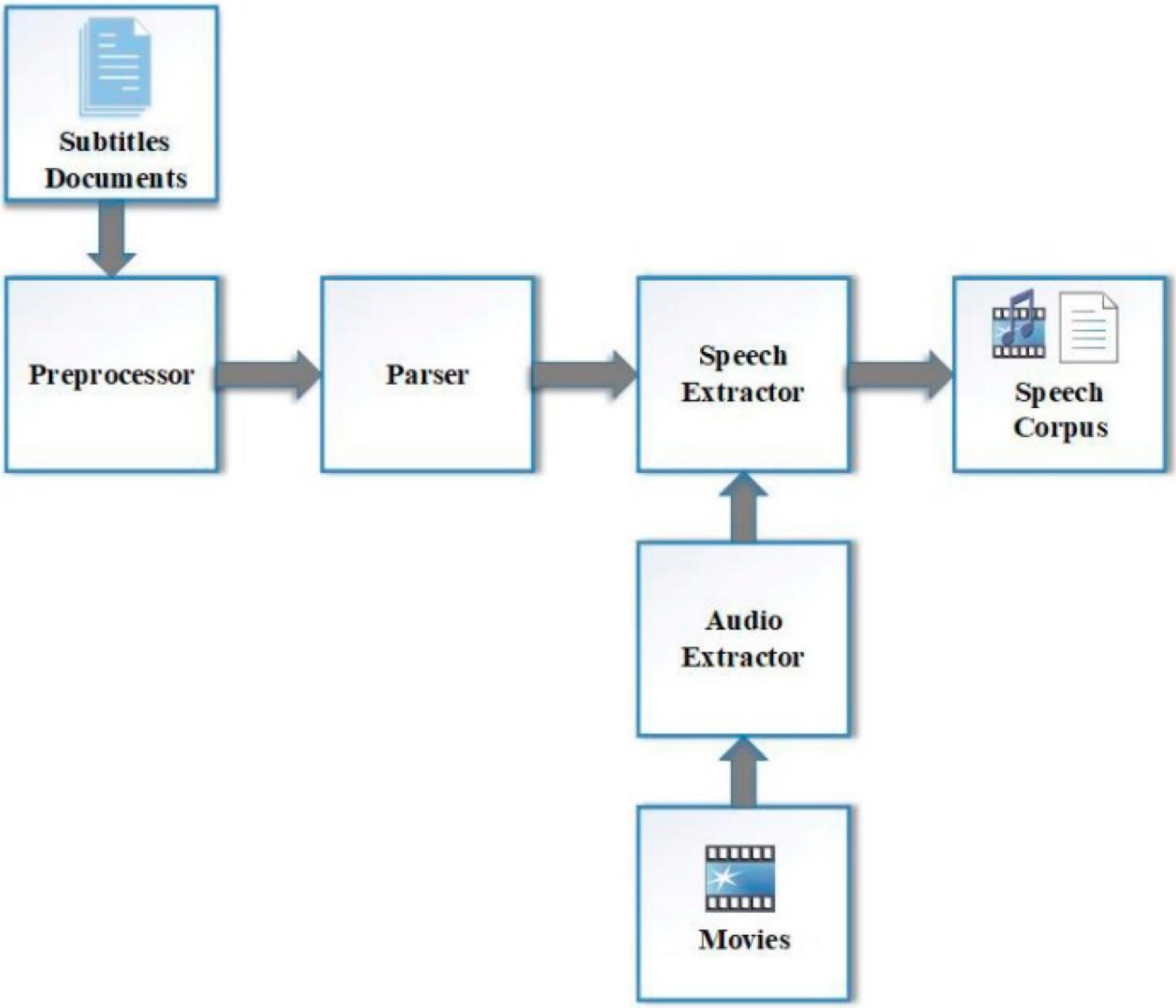

Figure 2: Speech recognition process
(Source: Barker et al. 2017, p.292)

The DTW are said to be superseded and the technique is processed according to dynamic time warping. Automated speech recognition can be specified to the well-known application that initiates the capacity to handle variant speeds of speeches. In a generalized way, an optimal match can be searched through the specified method. It permits the given sequences time period to be predicted by its computer to search all its allocated series where the restrictions are navigated (Barker et al. 2017). All the non-linear sequences are matched up as the sequences to be determined as 'warped'. The context composed of hidden Markov models can be utilized in the sequence of alignment method.

6. AUTOMATED SPEECH RECOGNITION COVERS END TO END PROCESSES

Guglani and Mishra (2018) stated that end-to-end processes that covers automatic speech recognition or ASR is said to be maintained since 2014. Research are said to be much more productive and effective in this fields. HMM-based approach is successfully working. Time or speed invests two sequences for which dynamic time warping can be signified as an algorithm towards the automatic speech recognition or ASR In the field of telephony, automatic speech recognition or ASR can keep its respected space. In the field of simulation and computer gaming system, it can be effective and wide informed. By the integration done with the IVR systems, it is seen that telephony systems can generate ASR or automatic speech recognition (Lojka et al. 2018). It is frequently fruitful to the contact centers. In spite of that, in a generalized manner personal computing can be integrated as the high level word processor. ASR or automatic speech recognition can be observed for the raised enhancement seen in the document production that can be notified with the coming days. Smartphones becomes efficient for improvising the speed of mobile processor where the efficiency can be seen for the recognition of the speech. User interface are utilized for speech developer that are most probably in use. Custom speech commands or predefined creation are the part of creation of speech recognition.

In the case of speech recognition, language learning initiates its utilities as its learning sources predicted by the language. Along with its speaking skills, it can provide the efficiency towards observing the appropriate command on pronunciation. It can develop the fluency to its speech additionally and can guide each individual (Helmke et al. 2018). Speech recognition can be benefitted to those students who can't see because of their blindness and they want to complete their education. Because of their low vision they are unable to read properly and want someone to help them so speech recognition is the best concept that becomes fruitful to them. Conveying words to the computer or commanding something by one's voice can initially transform the command to input all the words that can be helpful for them. By this way it makes them to serve themselves with the best result without looking into the keyboard the computer gives the output of the words in return that helps them to study in the appropriate manner (Greibus et al. 2017).

7. CONCLUSION

Speech recognition can be reflected in the computer that plays an important role for this process. Here, all the translated words can be acknowledged by the texts also. ASR or automated speech recognition can be recognized as speech to text 'STT' or computer speech recognition. Identification of the speaker voice can be recognized by their tone of the speech that can be referred to the speaker identification even the voice recognition. It is essential for the security process that it should be initiated to recognize the speaker voice. It can be fruitful towards simplifying the work as it can be translated according to the speech translation system. Speech recognition concept can be suggested for new variation of technologies as well as methodologies generated as interdisciplinary concept. fields, linguistics, computer science fields all are responsible for incorporating the knowledge and research that are linked with speech recognition. It measures the specific similarities. Observation can be detected by the person's movement. Instantly if the walking pattern of an individual is detected in a video that initiates accelerations and deceleration occurs due to the fast and slow movement analyzed by the walking patterns of a person. The intensity of capabilities is signified to be raised and seen in computers and the determined progress reports serves its positive value in this specified field. In early 1976, at the ending stage of the DARPA program, it was initiated to develop PDP-10 along with 4 MB ram by the researchers as the best computer. Along with its speaking skills, it can provide the efficiency towards observing the appropriate command on pronunciation. It can develop the fluency to its speech additionally and can guide each individual. 'Training' period can be signified as the requirement for the suggested kind of system analysis until then. In spite of that it can be subjected as the trainer to be suggestive for the children who can train their voices according to the gained response. Speech recognition can be benefitted to those students who can't see because of their blindness and they want to complete their education. In spite of that, in a generalized manner personal computing can be integrated as the high level word processor. ASR or automatic speech recognition can be observed for the raised enhancement seen in the document production that can be notified with the coming days.

REFERENCES

1. Barker, J., Watanabe, S., Vincent, E. and Trmal, J., 2018. The fifth'CHiME'speech separation and recognition challenge: dataset, task and baselines. arXiv preprint arXiv:1803.10609.
2. Fontan, L., Ferrané, I., Farinas, J., Pinquier, J., Tardieu, J., Magnen, C., Gaillard, P., Aumont, X. and Füllgrabe, C., 2017. Automatic speech recognition predicts speech intelligibility and comprehension for listeners with simulated age-related hearing loss. Journal of Speech, Language, and Hearing Research, 60(9), pp.2394-2405.

3. Greibus, M., Ringelienė, Ž. and Telksnys, L., 2017, April. The phoneme set influence for Lithuanian speech commands recognition accuracy. In 2017 Open Conference of Electrical, Electronic and Information Sciences (eStream) (pp. 1-4). IEEE.
4. Guglani, J. and Mishra, A.N., 2018. Continuous Punjabi speech recognition model based on Kaldi ASR toolkit. International Journal of Speech Technology, 21(2), pp.211-216.
5. Helmke, H., Slotty, M., Poiger, M., Herrer, D.F., Ohneiser, O., Vink, N., Cerna, A., Hartikainen, P., Josefsson, B., Langr, D. and Lasheras, R.G., 2018, September. Ontology for transcription of ATC speech commands of SESAR 2020 solution PJ. 16-04. In 2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC) (pp. 1-10). IEEE.
6. Hsu, W.N., Tang, H. and Glass, J., 2018. Unsupervised adaptation with interpretable disentangled representations for distant conversational speech recognition. arXiv preprint arXiv:1806.04872.
7. Hsu, W.N., Zhang, Y. and Glass, J., 2017, December. Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation. In 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 16-23). IEEE.
8. Iter, D., Huang, J. and Jermann, M., 2017. Generating adversarial examples for speech recognition. Stanford Technical Report.
9. Lojka, M., Viszlay, P., Staš, J., Hládek, D. and Juhár, J., 2018, September. Slovak broadcast news speech recognition and transcription system. In International Conference on Network-Based Information Systems (pp. 385-394). Springer, Cham.
10. Salimbajevs, A. and Ikauniece, I., 2017. System for Speech Transcription and Post-Editing in Microsoft Word. In INTERSPEECH (pp. 825-826).
11. Schönherr, L., Kohls, K., Zeiler, S., Holz, T. and Kolossa, D., 2018. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. arXiv preprint arXiv:1808.05665.
12. Tjandra, A., Sakti, S. and Nakamura, S., 2017, December. Listening while speaking: Speech chain by deep learning. In 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 301-308). IEEE